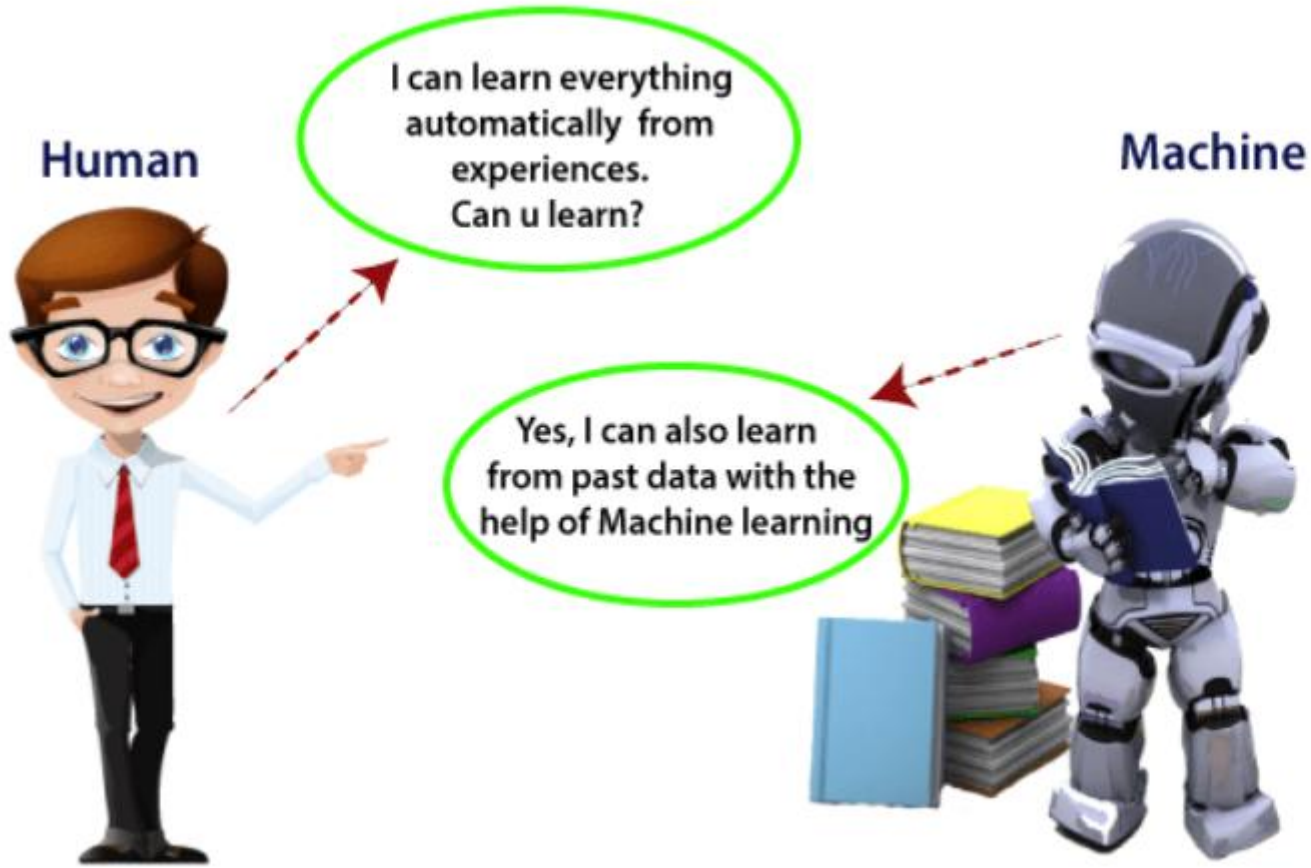
MACHINE LEARNING

What is Machine Learning?

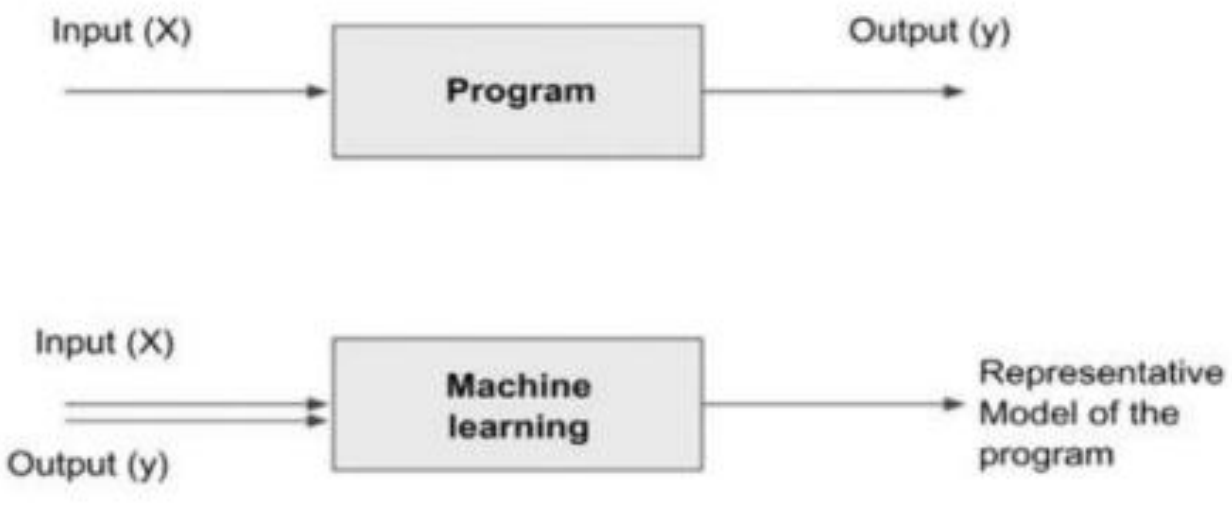
- Machine Learning (ML) is a branch of **Artificial Intelligence (AI)** that allows computers to **learn patterns from data** and make predictions or decisions **without being explicitly programmed(learns the rules/patterns by itself from data**, instead of a human writing the rules.).


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

Features of Machine Learning

1. **Learns from Data** – Models gain knowledge from past data and improve when more data is provided.
2. **Adapts Over Time** – Performance gets better as the model is trained with new or updated data.
3. **Finds Patterns** – ML detects hidden trends, relationships, and patterns that are not obvious.
4. **Handles Big Data** – Works effectively on large and complex datasets.
5. **Feature Selection** – Identifies the most important variables (features) that influence outcomes.
6. **Generalization** – Performs well on unseen/test data, not just on the training data.
7. **Supports Multiple Learning Types** –
8. **Supervised Learning** (learns from labeled data)
9. **Unsupervised Learning** (finds hidden patterns without labels)
10. **Reinforcement Learning** (learns from feedback/rewards).

Difference between program and ML



Need for Machine Learning

1. **Handling Large Data** – Traditional methods struggle with massive datasets, but ML efficiently processes and analyzes big data.
2. **Automation** – Reduces human effort by automating repetitive and complex tasks.
3. **Improved Accuracy** – ML models can make precise predictions and minimize human errors.
4. **Pattern Recognition** – Identifies hidden trends in data that are difficult for humans to detect.

5. **Cost Efficiency** – Reduces manual labor costs and increases productivity.
6. **Personalization** – Powers recommendation systems (e.g., Netflix, Amazon, YouTube).
7. **Real-Time Applications** – Used in self-driving cars, chatbots, and voice assistants.

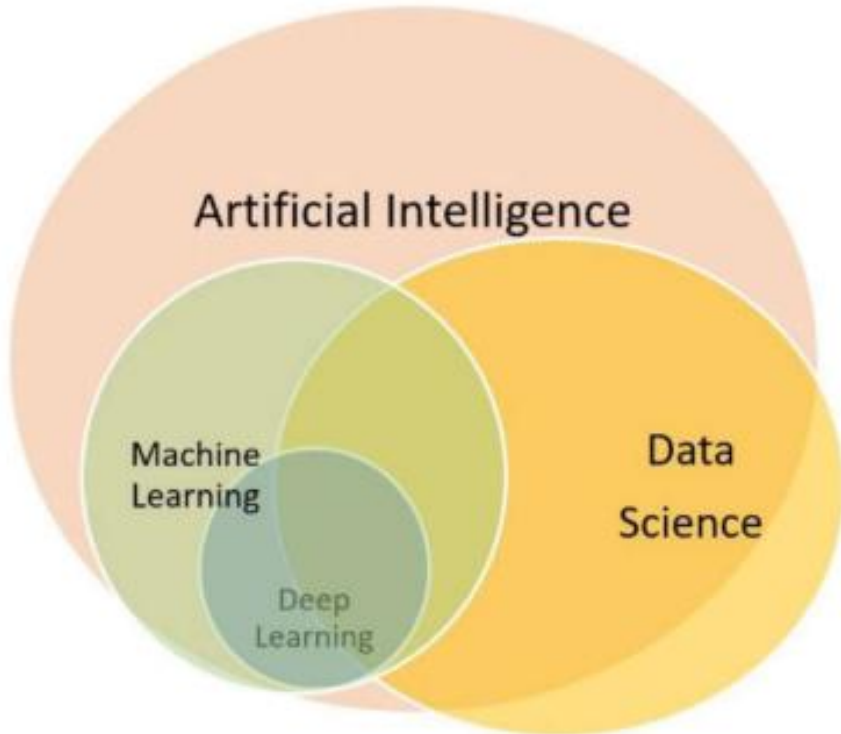
Applications of Machine Learning

1. Image recognition
2. Speech Recognition
3. Traffic prediction
4. Product recommendation
5. Self-Driving Cars
6. Automatic language translations
7. Healthcare
8. Agriculture
9. Finance

Difference between AI and ML

- **AI** is a field of computer science which makes a computer system that can mimic human intelligence
- 3 types
- 1.weak AI/narrow AI
- 2.General AI
- 3.Strong AI
- ML is a subfield of AI which enables machines to learn from past data or experience

Aspect	Artificial Intelligence (AI)	Machine Learning (ML)
Definition	AI is a broad field of computer science that makes machines **think, act, and perform tasks intelligently** like humans.	ML is a **subset of AI** that enables machines to **learn from data** and improve performance without explicit programming.
Goal	To create systems that can **simulate human intelligence** (reasoning, problem-solving, decision-making).	To **learn from data** and make predictions or decisions.
Scope	Wider concept that includes ML, Deep Learning, Robotics, NLP, Expert Systems, etc.	Narrower concept, mainly focused on algorithms that improve automatically through experience.
Approach	Focuses on **decision-making, logic, reasoning, and knowledge representation**.	Focuses on **data-driven learning and pattern recognition**.
Example	- Siri or Alexa understanding and responding to queries.	Self-driving cars. | - Netflix recommending movies. Email spam filter.


Artificial Intelligence
Machine Learning
Deep Learning
Data Science

DATASETS

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

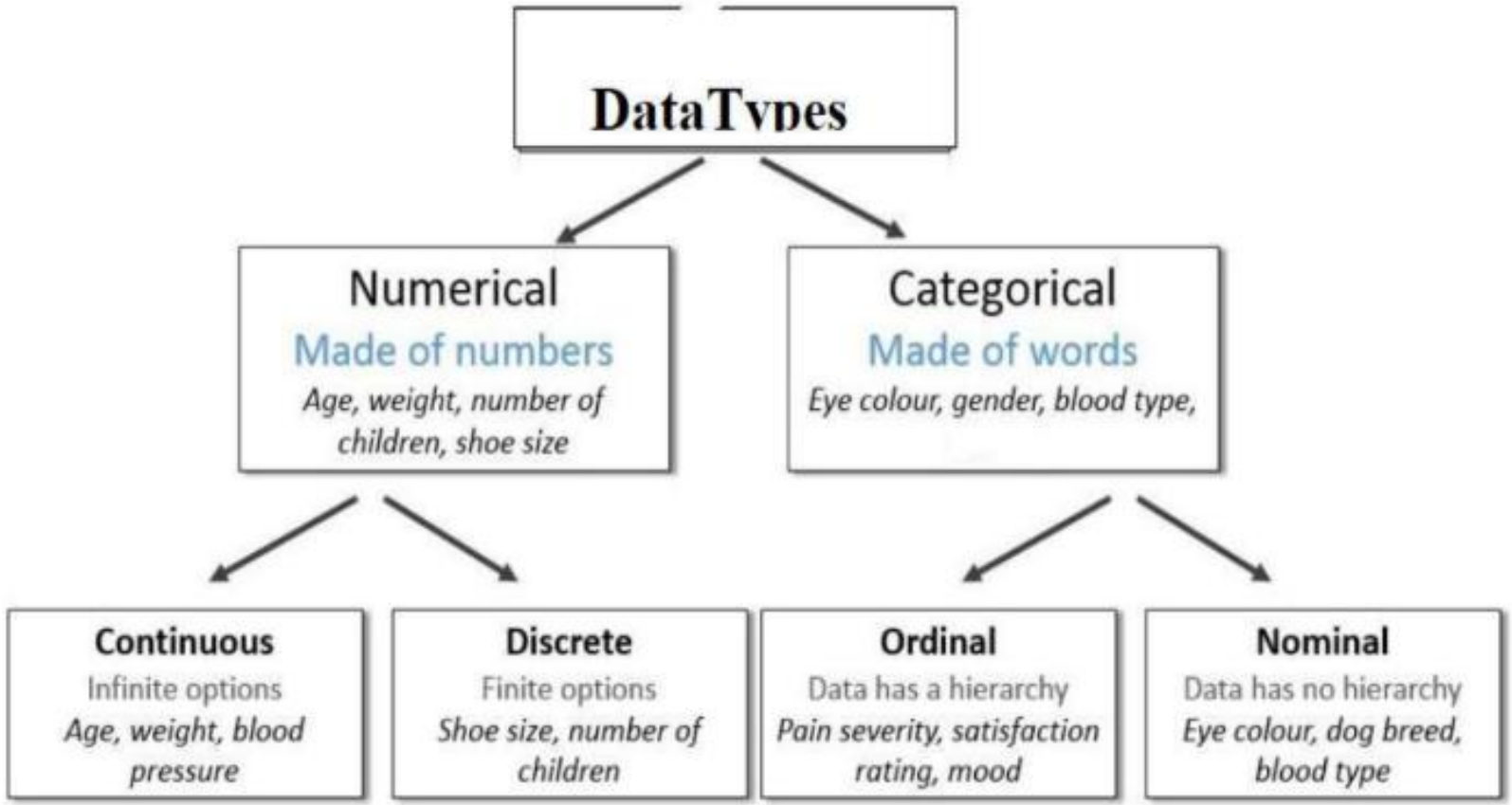
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1

- A dataset is a collection of data with a defined structure.
- This structure is also sometimes referred to as a "data frame".
- A data point (record, object) is a single instance in the dataset.
- Each instance contains the same structure as the dataset.

- An attribute (feature, input, dimension, variable, or predictor) is a single property of the dataset.

- Attributes can be numeric, categorical, date-time, text, or Boolean data types.

- A label (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes.


DataTypes
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type,
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

Popular Sources for ML Datasets

1. Kaggle dataset
2. UCI ML repository
3. Google dataset search engine
4. Scikit learn dataset

Training and Testing Data

1. Training Data

- The portion of the dataset used to **teach the ML model**.
- The model **learns patterns, relationships, and rules** from this data.
- Example:
    - Feeding a model past exam scores (Hours Studied → Marks) so it learns how study time affects marks.

2. Testing Data

- The portion of the dataset used to **evaluate the model** after training.
- It checks how well the model can make predictions on **unseen data**.

Example:

After training, test the model with new student data it hasn't seen to check if predictions are correct.

Why Split Data?

- If you train and test on the **same data**, the model may just **memorize** (overfitting).
- Using separate training and testing sets ensures the model **generalizes** to new situations.

Typical Split

- **70% Training**
- **30% Testing**
  (Sometimes 80%-20% or with a Validation set: 70%-15%-15%)

How Does It Work

1.Problem Definition

- Identify the business or research problem.
- Define goals and success metrics (e.g., accuracy, revenue impact).
- Example: Predict customer churn for a telecom company.

2. Data Collection

- ML starts with data (numbers, text, images, audio, etc.).
- Gather relevant data from sources such as databases, sensors, APIs, or logs.
- Ensure enough quantity and quality for training.
- Example: If we want to predict house prices, we collect data like **house size, location, number of rooms, and price**.

2. Data Preprocessing (Cleaning)

- Clean the data (handle missing values, duplicates, outliers).
- Transform categorical/numerical features.
- Normalize/scale features if required.
- Split data into training, validation, and test sets.

3.Exploratory Data Analysis (EDA)

- Understand patterns, distributions, and correlations.
- Use visualizations (histograms, scatter plots, heatmaps).
- Example: See which factors (like contract type or data usage) influence churn.

4. Feature Selection/Engineering

- Identify the **important variables (features)** that affect the output.
- Select the most relevant features to avoid overfitting and reduce complexity.
- Example: House price is affected by **size, location, age**, but not by **paint color**.

5. Splitting Data

- Data is divided into:
  - **Training Set (70–80%)** → Used to teach the model.
  - **Testing Set (20–30%)** → Used to check how well the model learned.

6. Choose a Model (Algorithm)

- Depending on the problem:
  - Predicting values → **Regression** (Linear Regression, SVR)
  - Classifying categories → **Classification** (Logistic Regression, SVM, Random Forest)
  - Grouping data → **Clustering** (K-Means, DBSCAN)

7. Training the Model

- The model learns patterns by adjusting its **parameters**.
- Feed training data to the chosen algorithm.
- Tune hyperparameters for better performance.
- The computer minimizes **error (difference between predicted & actual values)**.

8. Testing & Evaluation

- We test the model on unseen data.
- Measure performance using metrics like:
  - **Accuracy** (for classification)
  - **Mean Squared Error** (for regression)
  - **Precision, Recall, F1-Score** (for imbalanced data).

9. Prediction & Deployment

- Once the model performs well, it is deployed to make real-world predictions.
- Example: A trained spam filter labels new emails as *Spam* or *Not Spam*.

Machine learning models are broadly classified into **two main types**:

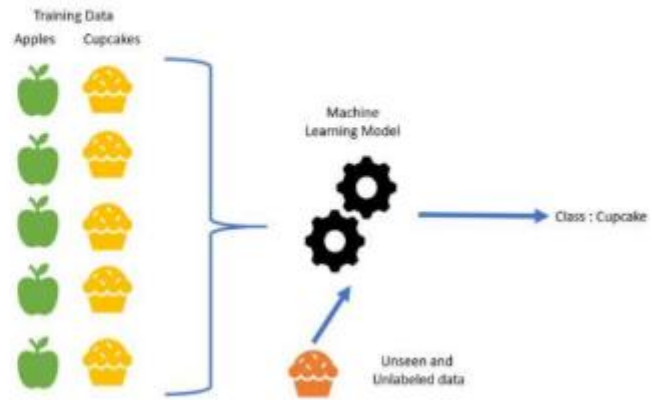
1. Supervised Learning
2. Unsupervised Learning

1. Supervised Learning

- Supervised learning involves training a model on **labeled data**, where each input is associated with a corresponding output.
- The goal is for the model to learn the relationship between inputs and outputs so that it can make accurate predictions on new, unseen data.

Input : Labeled Data

X $(features)$	Y $(labels)$
$x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots\ldots\ldots x_{1n}$	y_1
.	.
.	.
.	.
$x_{k1}, xk_2, x_{k3}, \ldots\ldots\ldots\ldots\ldots x_{kn}$	y_k



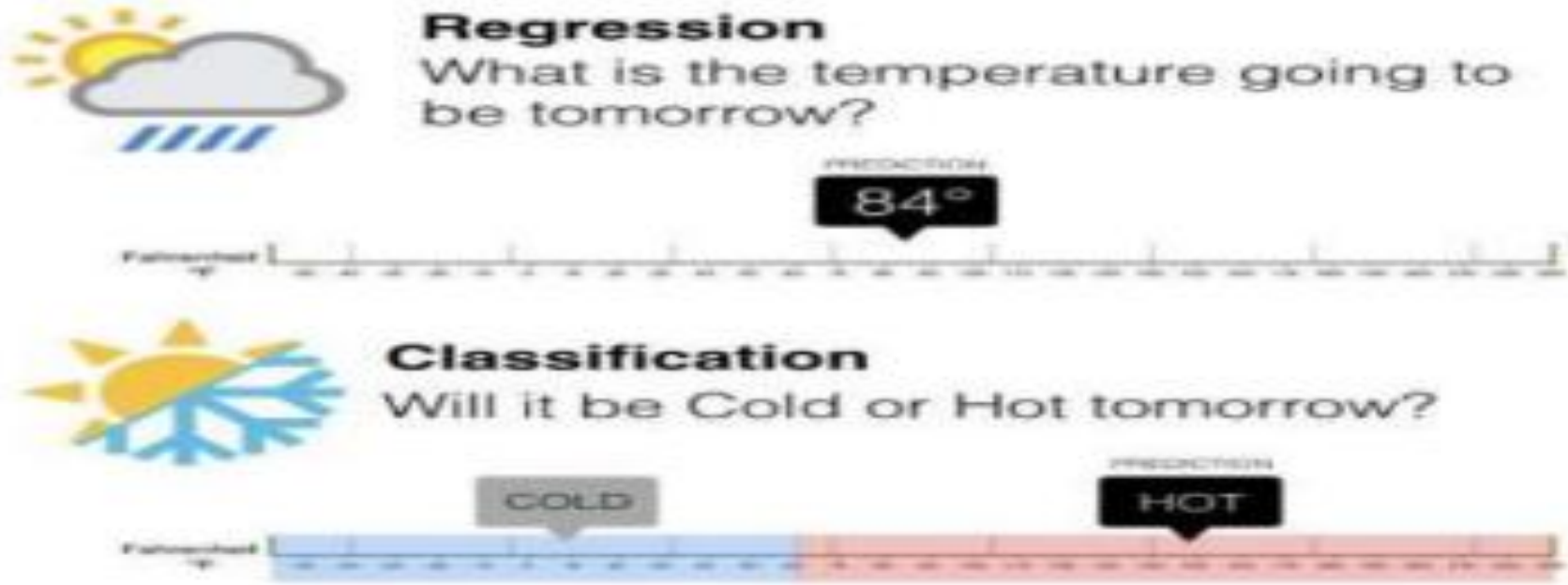
TYPES OF SUPERVISED LEARNING

1.Classification

- The output variable (target) is categorical (e.g., "spam" or "not spam").
- The model learns to assign new data points to predefined categories.
- **Examples:**
  - Email spam detection (Spam or Not Spam)
  - Handwritten digit recognition (digits 0-9)
  - Disease diagnosis (Healthy or Diseased)

2.Regression

- The output variable (target) is continuous (e.g., predicting a house price).
- The model learns to estimate numerical values based on input features.
- **Examples:**
  - Predicting house prices based on area, location, etc.
  - Forecasting stock prices.
  - Estimating temperature based on weather conditions.


Regression
What is the temperature going to be tomorrow?
PREDICTION
84°
Fahrenheit
Classification
Will it be Cold or Hot tomorrow?
COLD
PREDICTION
HOT
Fahrenheit

2. Unsupervised Learning

- Unsupervised learning deals with **unlabeled data**, meaning there is no predefined output. The model tries to learn the underlying patterns and structures in the data.
- **Types of Unsupervised Learning**
- **Clustering**
  - Groups similar data points together without predefined labels.
  - **Examples:**
    - Customer segmentation in marketing (grouping customers based on behavior).
    - Market basket analysis

Clustering is the process of identifying the natural groupings in a dataset.